DOCUMENT RESUME

ED 451 257 TM 032 485

AUTHOR Lee, Guemin; Fitzpatrick, Anne R.; Ito, Kyoko
TITLE Estimating Standard Errors for School PAAC's in Generalizability Theory.
PUB DATE 2001-04-14
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Cutting Scores; Elementary Education; *Elementary School Students; English; *Error of Measurement; Estimation (Mathematics); *Generalizability Theory; *Junior High School Students; Language Arts; Mathematics Tests; Performance Factors

ABSTRACT

School test performance is commonly summarized in terms of the percentage of students at or above a cut score (PAAC) that has been set on a test. Two approaches to estimating the standard errors for school PAACs were examined in this study: conditional standard errors and overall standard errors. The tests used were English language arts and mathematics tests administered in 1999 to students in grades 4 and 8 as part of a large statewide assessment. About 150 schools were randomly selected for the analyses. The results indicate that: (1) the conditional standard error appears to follow a quadratic pattern as a function of PAAC regardless of school size; (2) the quadratic shape is substantial when school size is small; and (3) there is distinct similarity between overall and conditional standard errors when they are conditioned on school size. Several feasible ways of reporting standard error information for school PAACs are also presented. (Contains 6 tables, 6 figures, and 13 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Guemin Lee

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 451 257

Estimating Standard Errors for School PAAC's In Generalizability Theory

Guemin Lee
Anne R. Fitzpatrick
Kyoko Ito
CTB/McGraw-Hill

TM032485

Paper Presented at the Annual Meeting
of the American Educational Research Association
Seattle, WA
April 14, 2001

BEST COPY AVAILABLE

Abstract

School test performance is commonly summarized in terms of the *p*ercentage of students *a*t or *a*bove a *c*utscore (PAAC) that has been set on a test. Two approaches to estimating the standard errors for school PAAC's were examined in this study: conditional standard errors and overall standard errors. The tests used in this study were English Language Arts and Mathematics tests administered in 1999 to Grades 4 and 8 students as part of a large, statewide assessment. About 150 schools were randomly selected for the analyses. The results indicated that (1) the conditional standard error appears to follow a quadratic pattern as a function of PAAC rgardless of school size, (2) the quadratic shape is substantial when school size is small, and (3) there is distinct similarity between overall and conditional SE's when they are conditioned on school size. Several feasible ways of reporting standard error information for school PAAC are also presented.

Estimating Standard Errors for School PAAC's In Generalizability Theory

Today it is common for school test performance to be described in terms of the *p*ercentage of students *a*t or *a*bove a *c*utscore (PAAC) that has been set on a test. That is, one or more proficiency levels have been set on the test, and the results of the assessments are reported in terms of the percentages of students who meet or exceed each proficiency level.

For several practical reasons, Cronbach, Bradburn, and Horvitz (1994) recommended that results relative to one proficiency level be used when schools and/or districts are compared in terms of their PAAC's. To do this, students in a school are assigned one of two classifications on the basis of their test performance. Specifically they are classified as performing "below a particular cutscore" or "at or above cut score." The student results are often aggregated at the school or district level to produce an overall PAAC for the school or district.

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) included a recommendation of reporting conditional standard errors:

> Standard 2.14: Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. (p. 35)

Overall standard errors as well as conditional standard errors are recommended reported together:

> Standard 2.2: The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation. (p. 31)

Although schools and districts routinely are evaluated in terms of PAAC's, there currently are only a few studies that have considered how the practitioner should estimate and report the standard errors of these PAAC's. At least, two approaches exist for estimating standard errors for school PAAC's (Linn & Burton, 1994; Yen, 1997). Both methods provided procedures for estimating standard errors for school PAAC's using variance component estimates, but the approaches are different. The procedures described in the current study is more aligned to Yen's method in that both procedures dichotomize student scores into

pass/fail status and uses variance component estimates to estimate standard errors for school PAAC's. However, the current study uses a different algorithm to estimate the variance components and it follows a more typical generalizability theory framework than Yen's procedures. This study also applied conditional and overall standard errors to the same data sets and investigated the properties of both types of standard errors.

The specific objectives of the current study were to

1. explore the properties of overall and conditional standard errors for school PAAC's,
2. determine the effect of school size on both the overall and conditional standard errors for school PAAC's,
3. evaluate the relative appropriateness of using the overall and conditional standard errors for school PAAC's, and
4. suggest some practical ways to report standard error information for school PAAC's.

Procedures of Estimating Standard Error for School PAAC

The conditional approach to estimating standard errors (SE's) produces results that vary with the specific value of PAAC. The overall approach to estimating SE's does not. However, by either approach, the SE estimates produced are dependent upon the number of students who take the test.

Conditional Standard Error Approach

The classification of students "at or above cut score" or "below cut score" involves scoring students dichotomously, as 0 or 1, to reflect their status relative to a cut score. The vector of students' status scores in a school can be thought of as independent binomial trials centered on a certain proportion for that school. For binomial variables, the amount of error associated with a specific proportion is expected to vary as a function of the PAAC.

Let a dummy variable, X_{ps}, denote the status score for student p within school s, where each student's reaching/not-reaching status is defined,

$$X_{ps} = \mu + \mu_s \sim + \mu_{p:s,e} \sim . \qquad (1)$$

The terms of right-hand side are grand mean, school effect, and person within school effect confounded with unexplained sources of error, respectively. The PAAC for school *s* can be estimated by

$$\hat{PAAC}_s = \frac{\sum_{p=1}^{n_s} X_{ps}}{n_s} \times 100, \quad (2)$$

where n_s is the number of students who took a test in the school *s*.

Under a generalizability theory framework, the absolute error for the PAAC of school *s* is defined as

$$\Delta_s = \hat{PAAC}_s - PAAC_s, \quad (3)$$

where $\hat{PAAC}_s$ is an estimate of the PAAC for school *s* over a sample of students, and $PAAC_s$ is the true PAAC for school *s* over infinite population of students. The variance for this error score for school *s* is

$$\sigma^2(\Delta_s) = Var(\hat{PAAC}_s - PAAC_s). \quad (4)$$

Because $PAAC_s$ is a constant, Equation 4 should be the same as

$$\sigma^2(\Delta_s) = Var(\hat{PAAC}_s). \quad (5)$$

By the central limit theorem (Hogg & Craig, 1995),

$$\sigma^2(\Delta_s) = Var(100 \times \frac{\sum_{p=1}^{n_s} X_{ps}}{n_s}) = \frac{100^2 \times Var(X_{ps})}{n_s}, \text{ where } Var(X_{ps}) = \frac{\sum_{p=1}^{n_s} (X_{ps} - \bar{X}_s)^2}{n_s - 1}. \quad (6)$$

Therefore, an estimator of the standard error for PAAC of school *s* is

$$SE(PAAC_s) = 100 \times \sqrt{\frac{\sum_{p=1}^{n_s} (X_{ps} - \bar{X}_s)^2}{n_s(n_s - 1)}}. \quad (7)$$

Overall Standard Error Approach

In the situation where schools administer a single test form, the univariate $p : s$ generalizability study design, persons (*p*) nested within schools (*s*), is appropriate for estimating variance components. The linear model for the response of a person within a school, which is the same as Equation 1, treats schools as objects of measurement and persons as a random facet.

To obtain an overall standard errors for school PAAC's, principles from generalizability theory were used to estimate variance components. Those represent the score variances for a single person within a school with a test form. The variance component estimates were then used to derive the overall estimate for a set of students within a school using the central limit theorem. The overall standard error was estimated by the following formula

$$SE(PAAC) = 100 \times \sqrt{\frac{\hat{\sigma}^2_{p,s}}{n'_s}}, \tag{8}$$

where $\hat{\sigma}^2_{p,s}$ is the estimator of generalizability study variance component for students nested within schools, and n'_s is the number of students in a decision study. The standard error for PAAC in Equation 8 is not dependent upon a specific PAAC. However, the standard error in Equation 7 varies with specific school PAAC's.

Method

Data Sources

The tests used in this study were English Languate Arts (ELA) and Mathematics (MA) tests administered in 1999 to students in Grades 4 and 8 as part of a large, statewide assessment. More than 3,600 schools and approximately 200,000 to 250,000 students per grade were involved in the assessment program. The general characteristics of each test are presented in Table 1.

Insert Table 1 About Here

Because both the ELA and the MA tests were composed of multiple-choice (MC) and constructed-response (CR) items, two item response models were used for the scaling, the three-parameter logistic model (Lord & Novick, 1968; Lord, 1980) and the two-parameter partial credit model (Muraki, 1992; Yen, 1993). Approximately 2,000 to 7000 students were randomly sampled from tested students for calibration. The item parameters were estimated using the PARDUX computer application program (Burket, 1996).

The Bookmark standard setting procedure (Lewis, Mitzel, & Green, 1996) was implemented to set cutscores on the tests. Three cutscores were set in each grade to define four performance levels. Students in

a given performance level were expected to perform the majority of what is described for that level and even more of what is described for levels below. Performance levels and descriptions for the grade 4 Mathematics test are presented in Table 2 as an example.

Insert Table 2 About Here

In this study, Performance Level 3 was selected to compute PAAC for each school. Students were dichotomously classified in terms of their performance related to the cut score used to separate performance levels 2 and 3. About 150 schools were randomly selected for the analyses from schools that had tested at least two students. Summary statistics describing the schools in the sample for each grade and each content are presented in Table 3.

Insert Table 3 About Here

Analyses

For estimating conditional standard errors for school PAAC's, the individual school's PAAC, enrollment, and students' status scores were used as the inputs to Equation 7. A generalizability study (G study) analysis was conducted to estimate overall standard error. Because the number of students for each school varied, the conditions for a balanced design were not met in this situation. Accordingly, the urGENOVA computer application program (Brennan, 1999) was used to estimate the variance components for the unbalanced design. Following the G study, decision (D) studies were conducted to investigate the effects of school size on estimates of the overall standard errors for school PAAC's. The school sizes in D-studies varied by m_j *10, and m_j =1, 2, …, 35.

Results and Discussion

Conditional Standard Errors for School PAAC's

Figure 1 illustrates conditional standard errors (SE's) conditioning on school PAAC's for grades 4 and 8 ELA and MA tests. The horizontal axis represents estimated school PAAC's and vertical axis

represents associated conditional SE's. Schools with '0' or '100' PAAC's were not included in the plots because their SE's were zeros by definition.

Insert Figure 1 About Here

These plots show that the conditional SE's varied to some degree as a function of the size of PAAC. Because students within a school were dichotomously scored, the standard errors follow a concave-down quadratic function in general (Lord, 1955, 1957; Brennan, 1998). This quadratic trend among conditional SE's was relatively less clear than that found by Brennan (1998). One reason for this unclear pattern in the current study might be that the magnitudes of conditional SE's for school PAAC's are affected by two factors, the sample size of students and a specific PAAC of a school, as shown in Equation 7. The effects of these two factors are confounded in the plots.

To see the more direct relationship between PAAC's and conditional SE's, schools were divided into three groups: small-, middle-, and large-size schools. Small-size schools have less than 50 students, and large-size schools have more than 150 students. Other schools were classified as middle-size schools. Conditional SE's for school PAAC's for these small-, middle-, and large-size schools for grade 4 ELA test are presented in Figure 2.

Insert Figure 2 About Here

The clearer quadratic pattern among conditional SE's can be seen in the top graph of Figure 2 for the small-size schools. For middle- and large-size schools, conditional SE's seemed to have a rectangular distribution. The larger the student sample size is, the less clear the quadratic trend among conditional SE's is. That is, the quadratic pattern among the conditional SE's for school PAAC's seemed to be mitigated by the school size. The pattern in conditional SE's was not clear due to the confounded effects of school size with school PAAC.

One note should be mentioned at this point. The vertical axes for three plots in Figure 2 were scaled from 0 to 30 for allowing relative comparison across three-size schools. Thus, it was difficult to find

quadratic patterns in middle- and large-size schools via eyeball tests. However, this does not mean there was no quadratic pattern among estimated conditional SE's for these schools. In Figure 3, conditional SE's for school PAAC's for the middle-size schools are re-presented on a re-scaled vertical axis from 0 to 7. Now, we can see the clearer quadratic pattern. Because most conditional SE's for school PAAC's for middle-size schools belonged to a band of 4% to 6%, we got an impression of the rectangular distribution in Figure 2 plots.

Insert Figure 3 About Here

The relationship between student sample size and conditional SE's for school PAAC's is shown in Figure 4. Very clear patterns can be found across four plots for different grades and content areas. Conditional SE's for school PAAC's decreased as the student sample size increased. However, after student sample size of 150, similar conditional SE's were reported. As indicated previously, the magnitude of conditional SE for a school depends upon both the student sample size and a specific PAAC. We had difficulty in observing patterns in Figure 1 plots, but we can see clear patterns in Figure 4 plots. This suggests that the conditional SE's for school PAAC's were more strongly related to the student sample size than to the specific PAAC's.

Insert Figure 4 About Here

Overall Standard Errors for School PAAC's

Table 4 shows variance component estimates from the G-study for the random effects *p:s* design for each grade and each content.

Insert Table 4 About Here

The school variance is an estimate of the variance of school student-status mean scores. This variance component serves as a universe-score variance (analogous to true score variance in the classical

test theory) because schools are the object of measurement in this situation. Table 4 shows that, over grades and subject areas, only about 15-20% of the total variance was accounted for by schools and most of the variance was attributable to pupils and unexplained sources of errors.

Overall SE's for school PAAC's were estimated using Equation 8. Overall SE's centered upon student sample size are presented in Figure 5. Overall SE's for school PAAC's depend upon only student sample size, but not upon specific PAAC's. Although the overall SE's decreased as the student sample size increased, after student sample size of 150, the overall SE's did not change.

Insert Figure 5 About Here

To check similarity between conditional and overall SE's conditioning on student sample size, the plots in Figures 4 and 5 for grade 4 ELA are plotted together in Figure 6. Empty circles represent conditional SE's and solid diamonds represent overall SE's. Close relationship between overall and conditional SE's can be observed.

Insert Figure 6 About Here

Reporting Standard Errors for School PAAC's

The results of this study can be summarized by the three major findings: [1] rgardless of school size, the conditional SE appears to follow a quadratic patterr as a function of PAAC with a peak at the middle of PAAC range, [2] the quadratic shape is substantially pronounced when school size is small, [3] relatively constant SE values are expected for middle- or large-size schools regardless of school PAAC, and [4] the overall and conditional SE's are similar, when they are conditioned on student sample size. Based upon these major findings, two approaches to reporting the overall and conditional SE's for school PAAC's are suggested below.

Presenting overall standard errors for school PAAC's and the corresponding student sample sizes would be one good way in practice providing standard error information. A table suggested for test developers to report overall SE's based upon different student sample sizes is given in Table 5.

Insert Table 5 About Here

To provide information on conditional SE's, separate SE table should be produced for groups of schools that differ in size. In the following example, we classified schools into three different sizes, and presented conditional SE's for school PAAC's for each size group separately.

Insert Table 6 About Here

The conditional SE's in Table 6 are **fitted** estimates using polynominal regression of degrees 2. Brennan (1998, p. 315) provided two reasons for using fitted conditional SE's rather than unfitted observed estimates in reporting standard error information:

> "1) In most testing programs, it is difficult, and often unacceptable, to treat examinees receiving the same score differently; and examinees with similar scores expect to be treated similarly.
> 2) The obtained results are subject to random sampling error; and error introduced by using fitted values may be considerably less than sampling error."

In our example, for large-size schools, 3% SE can be applied to schools with very low (less than 15) or high (greater than 80) PAAC's. To other PAAC's, a 4% SE could be used. An alternative would be to use a 4% SE for all school regardless of PAAC's. In this case, the use of one SE is not likely to lead to any significant misinterpretation about school PAAC's. In a similar manner, we might consider the use of 5% SE for all middle-size schools. However, for small-size schools, there are non-negligible differences among conditional SE's based upon different PAAC's. Thus, it would be necessary to report different conditional SE's for different PAAC's in small-size schools.

The simultaneous use of overall and conditional SE's for school PAAC's can be considered. That is, if student sample size is greater than 50, apply one SE for schools regardless of their specific PAAC's (e.g., apply 5% to schools with students greater than 50 and less than 150, and 4% to school with students

greater than 150). If student sample size is less than 50, conditional SE's can be used as shown in the "Small" column of Table 6.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Brennan, R.L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. Applied Psychological Measurement, 22, 307-331.

Brennan, R.L. (1999). Manual for urGENOVA (version 1.4). (Iowa Testing Programs Occasional Paper No. 46). Iowa City, IA: University of Iowa.

Bucket, G.R. (1996). PARDUX [Computer program]. Monterey, CA: CTB/McGraw-Hill.

Cronbach, L.J., Bradburn, N.M., & Horvitz, D.G. (1994). Report of the Select Committee: Sampling and statistical procedures used in the California Learning Assessment System.

Hogg, R.V., & Craig, A.T. (1995). Introduction to mathematical statistics. Englewood Cliffs, NJ: Prentice Hall.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996, June). Standard setting: A Bookmark approach. In D.R. Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Lord, F.M. (1955). Estimating test reliability. Educational and Psychological Measurement, 15, 325-336.

Lord, F.M. (1957). Do tests of the same length have the same standard error of measurement? Educational and Psychological Measurement, 17, 510-521.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Menlo Park, CA: Addilon-Wesley.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

BEST COPY AVAILABLE

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187-213.

TABLE 1
Descriptive Statistics for Tests Used in This Study

	English Language Arts		Mathematics	
	Grade 4	Grade 8	Grade 4	Grade 8
No. of Students Tested	198,785	207,035	245,088	218,448
No. of Items	32	29	48	45
No. of MC items	28	25	30	27
No. of CR items	4	4	18	18
Total Score Points	42	43	70	69
Raw Score Mean	28.6	30.2	48.5	37.3
Raw Score S.D.	7.02	7.34	12.90	15.16

Note. MC = Multiple choice; CR = Constructed response.

TABLE 2

Grade 4 Mathematics Performance Levels and Descriptions

Performance Level	Descriptions Simplified for the Paper
4	Students use estimation, probabilistic prediction, and graphical representations, and identify equivalence and complex measures. They order decimals and identify, create, and describe combinations and patterns. They also analyze situations, apply and explain reasoning, and draw conclusions.
3	Students consistently understand probabilities, percents, and relationships among fractions, and identify patterns and parts of various figures. They work with and interpret real-world data. They also solve multi-step problems and present reasonable solutions with justifications.
2	Students generally are able to use all basic operations and demonstrate an understanding of whole-numbers. They use manipulatives to solve for an unknown and to model simple fractional relationships. They also identify various shapes and patterns and interpret data.
1	Students may use some of basic operations and show some understanding of simple concepts, data, and figures. They may use manipulatives to explore patterns and represent whole-number relationships.

TABLE 3

Descriptive Statistics for School PAAC's for Each Grade and Each Content

Grade/ Content	No. of Schools	No. of Students	No. of Students in a School			School PAAC		
			Mean	SD	Range	Mean	SD	Range*
Grade 4								
ELA	151	9,988	66	48.6	2 ~ 272	47.5	21.92	2.6 ~ 94.8
MA	156	11,187	72	59.6	2 ~ 284	64.5	24.00	2.7 ~ 97.5
Grade 8								
ELA	157	14,110	90	110.1	2 ~ 520	46.2	24.73	2.9 ~ 88.5
MA	155	16,136	104	119.0	2 ~ 532	37.2	24.54	1.9 ~ 90.0

Notes. ELA = English and Language Arts; MA = Mathematics; Range* for school PAAC does not include 0 or 100 PAAC's.

TABLE 4

Variance Component Estimates for the Random Effects Pupil (p) Nested within School (s) Effects

Variance Component	Grade 4		Grade 8	
	ELA	MA	ELA	MA
$\hat{\sigma}^2(s)$	3.5 (16.0%)	4.7 (20.3%)	3.2 (12.8%)	5.2 (21.7%)
$\hat{\sigma}^2(p:s)$	21.4 (84.0%)	18.4 (79.75)	21.8 (87.2%)	18.8 (78.3%)

Note. To compare comprehensively, the variance component estimates were multiplied by 100 and then rounding to one decimal place. ELA = English Language Arts; MA = Mathematics.

TABLE 5
Overall Standard Errors of Specific PAAC's for Grade 4 English/Language Arts Test
Based upon Student Sample Size

Student Sample Size	Standard Error
5	21
10	15
15	12
20	10
25	9
30	8
40	7
50	7
60	6
70	6
80	5
90	5
100	5
120	4
140	4
160	4
180 or More	3

TABLE 6

Conditional Standard Errors of Specific PAAC's for Small-, Middle-, and Large-Size Schools for Grade 4 English/Language Arts Test

School PAAC	School Size		
	Small	Middle	Large
5	3	3	3
10	5	4	3
15	6	4	3
20	7	5	4
25	8	5	4
30	9	5	4
35	10	5	4
40	10	6	4
45	11	6	4
50	11	6	4
55	11	6	4
60	11	6	4
65	11	5	4
70	11	5	4
75	10	5	4
80	10	5	3
85	9	4	3
90	8	4	3
95	7	3	3

Notes. Small-size schools = 2 $\leq$ No. of students < 50; Middle-size schools = 50 $\leq$ No. of Students < 150; Large-size schools = No. of students > 150. All standard error estimates were rounded to the first digit numbers.

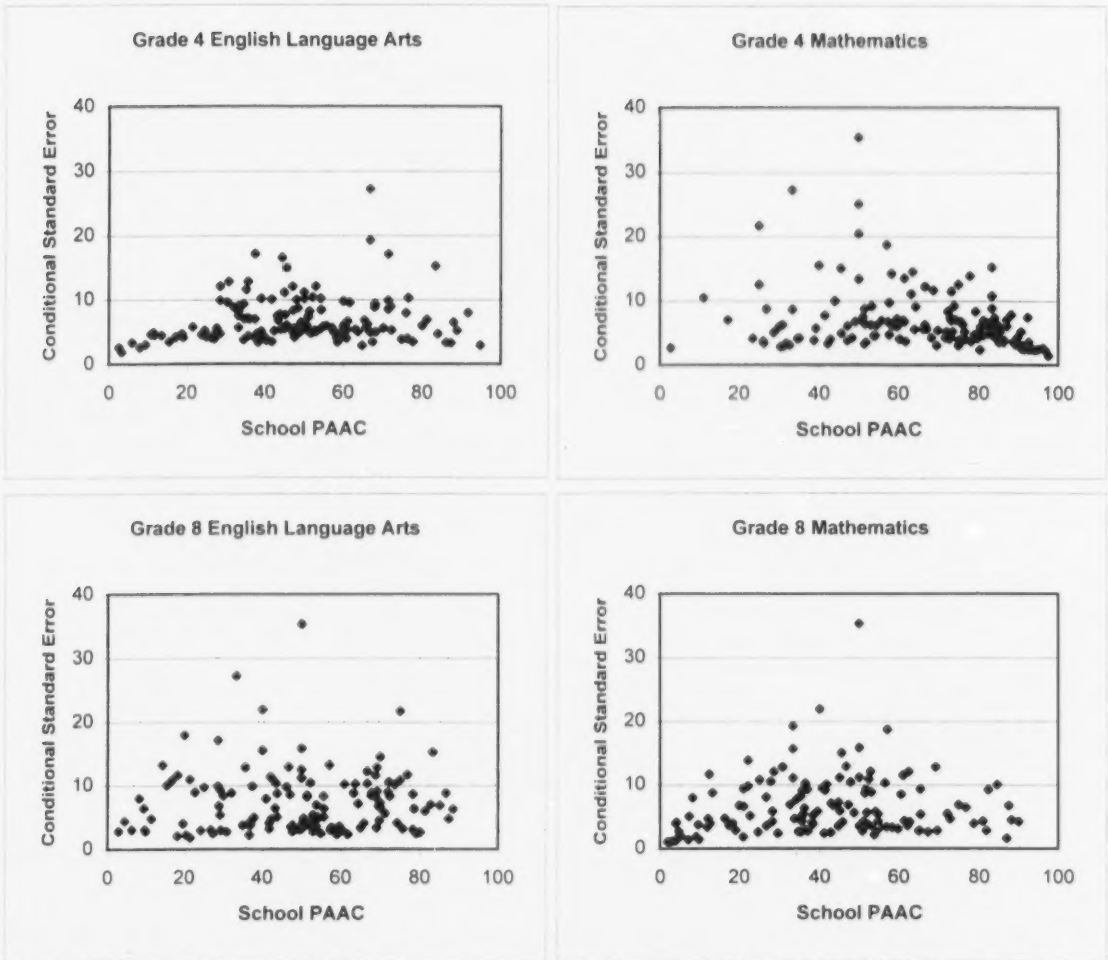

Figure 1. Conditional standard errors for school PAAC's for Grades 4 and 8 English Language Arts and Mathematics Tests

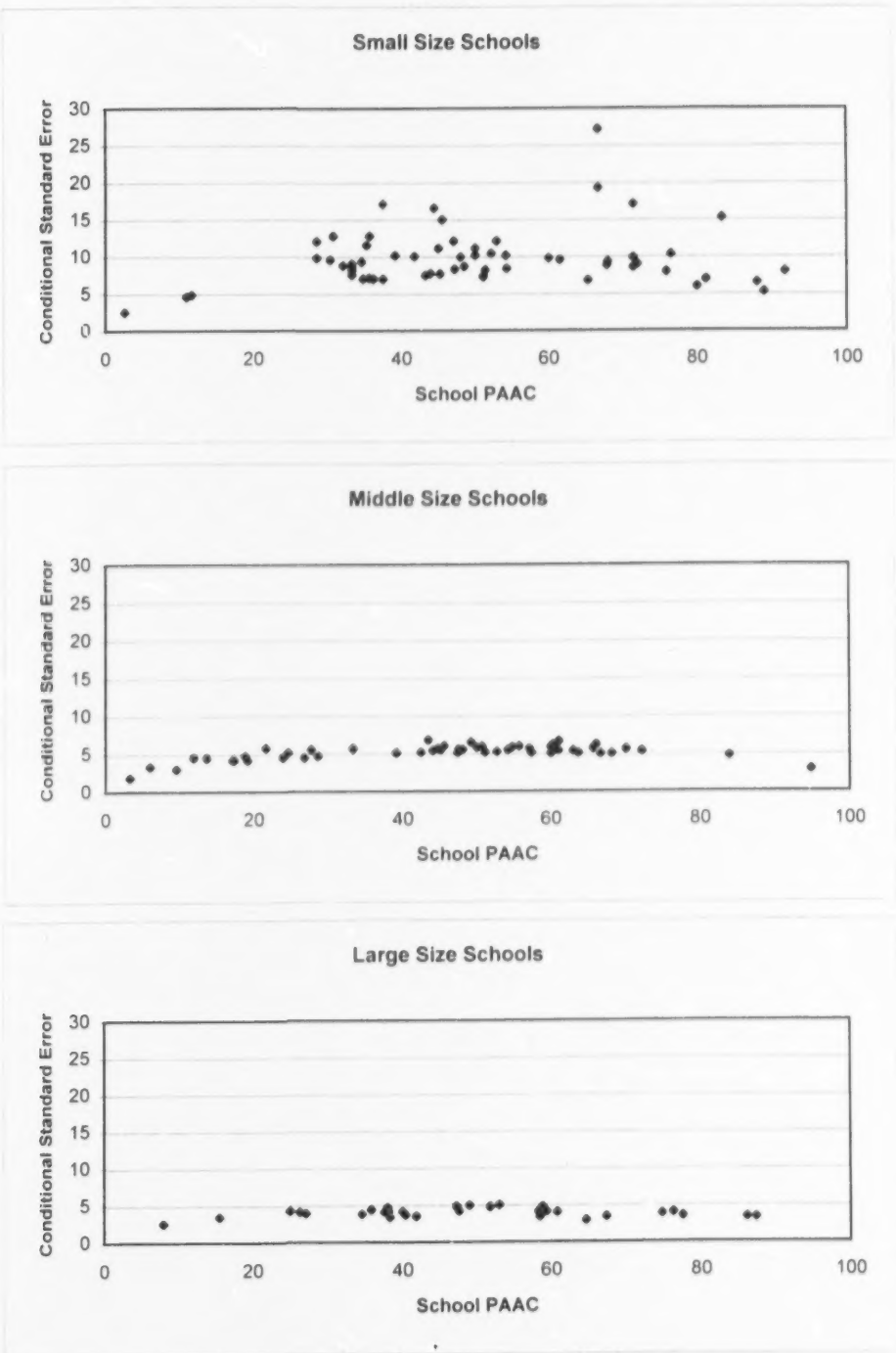

Figure 2. Conditional standard errors for school PAAC's for Grade 4 English Language Arts test for small-, middle-, and large-size schools.

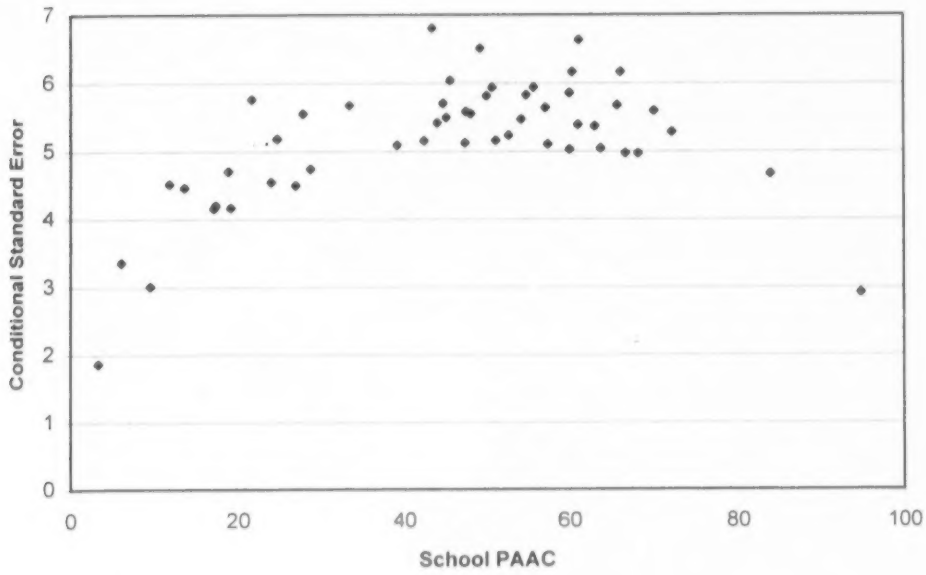

Figure 3. Conditinal standard errors for school PAAC's for Grade 4 English Language Arts test for middle-size schools.

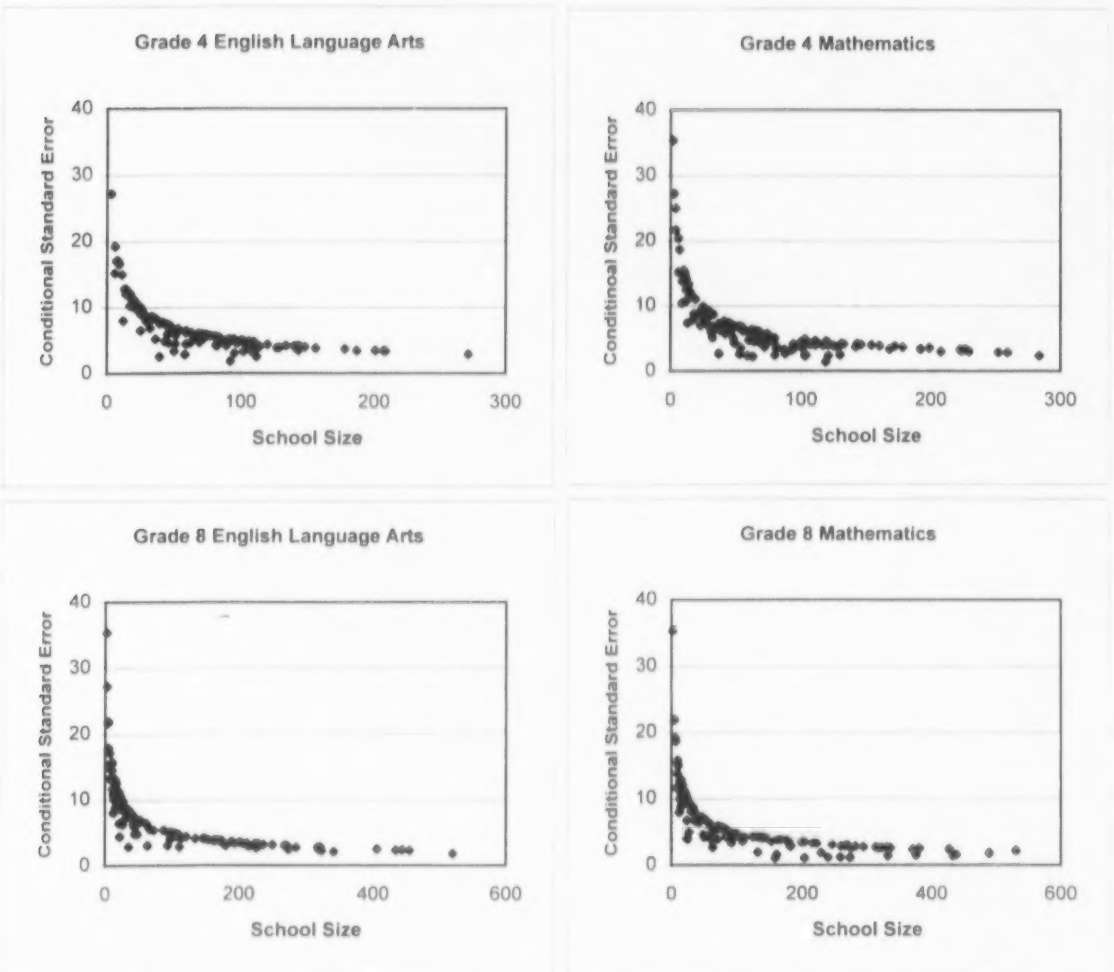

Figure 4. Conditional standard errors for school PAAC's for Grades 4 and 8 English Language Arts and Mathematics tests based upon school size.

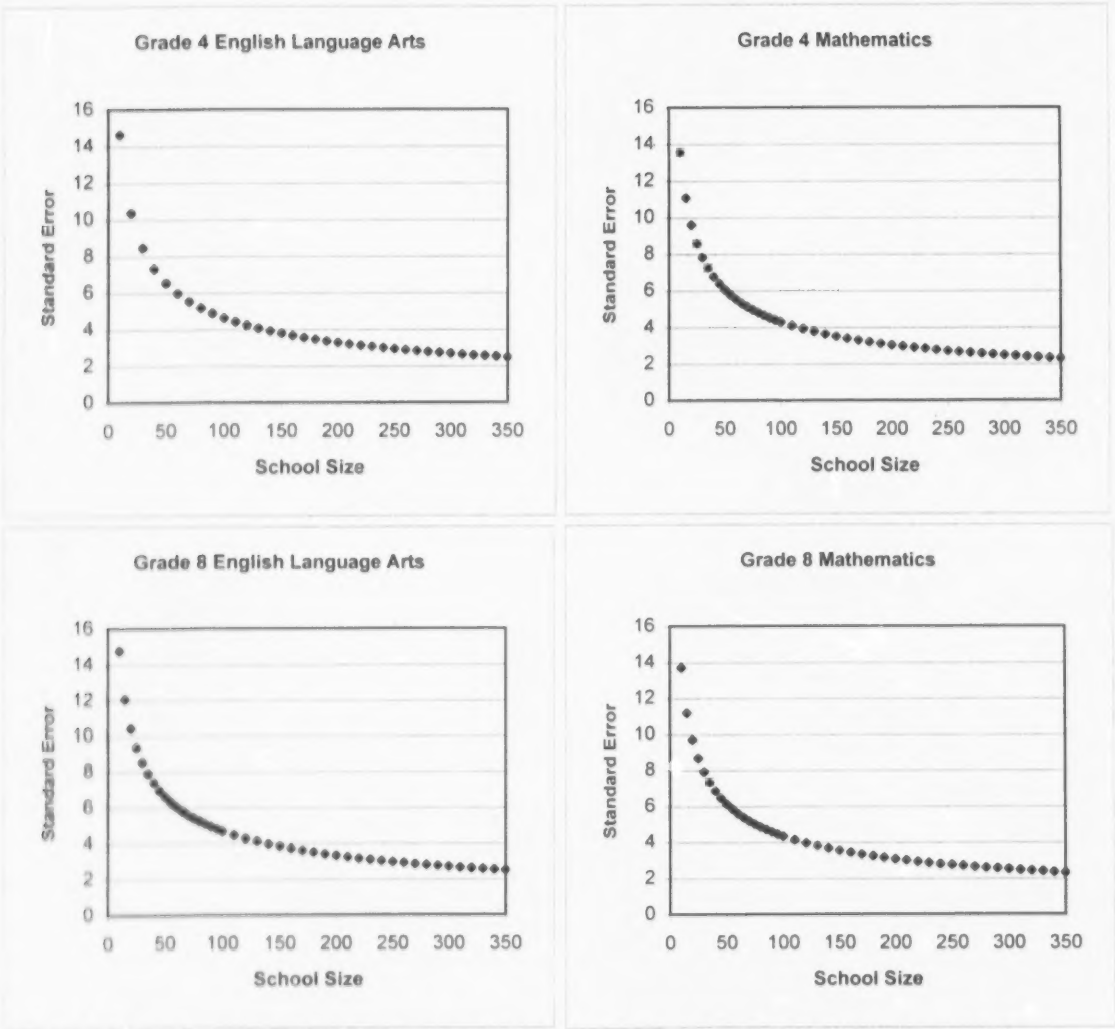

Figure 5. Overall standard errors for school PAAC's for Grades 4 and 8 English Language Arts and Mathematics tests based upon school size.

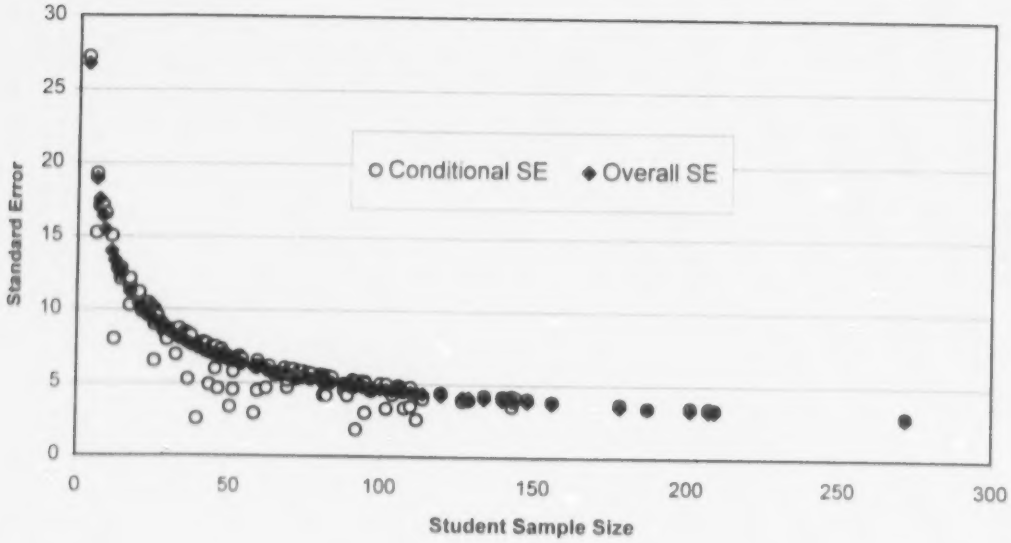

Figure 6. Overall and conditinoal standard errors for school PAAC's for Grade 4 English Language Arts test based upon student sample size.